# Sense Tagged CLE Urdu Digest Corpus

Saba Urooj, Sana Shams, Sarmad Hussain, Farah Adeeba
*Centre for Language Engineering, Al-Khawarizmi Institute of Compute Science,*
*University of Engineering and Technology, Lahore*
*firstname.lastname@kics.edu.pk*

## Abstract

*This paper presents the construction of an Urdu Sense Tagged corpus using four main lexical resources; an Urdu wordlist consisting of 5000 high frequency content words, a 100K words corpus annotated with part of speech (POS) tags, an Urdu WordNet with approximately 5058 senses and Urdu morphological analyzer. The paper also briefly presents Urdu word-sense annotation tool, a software tool developed to provide an easy interface for sense tagging, ensuring tagging consistency and accelerating the annotation speed. In this version of the Urdu sense tagged corpus, 17,006 words have been sense tagged with 2285 unique senses. The final section discusses the linguistic and tool specific challenges in the construction of sense tagged corpus and describes future work in this context.*

## 1. Introduction

Words possess multiple senses and each sense is context sensitive where sense can be defined as semantic value (content) of a word when compared to other words; i.e. when it is part of a group or set of related words. The process of assigning senses to a word is not simple as those senses are dissimilar and in some cases entirely distinctive [1]. Therefore, most of the senses of words are clearly different in meaning as the distinction between paper as 'newspaper' and paper as 'a material used for writing' while others are not as obvious. For example[1], paper as newspaper can have following related senses: a daily or weekly publication on folded sheets containing news and articles and advertisements, a business firm that publishes newspapers, the physical object that is the product of a newspaper publisher. Similarly in Urdu language, "پرچہ" (pərʧɑ:\paper) can have following related senses: "امتحان کے سوالات کا کاغذ" (ɪmʈəhɑ:n ke: səvɑ:lɑ:ʈ kɑ: kɑ:ɣəz/exam paper), " کاغذ کا ٹکڑا" (kɑ:ɣəz kɑ: ʈukɽɑ:/piece of paper) and " ہفتہ وار اخبار/ رسالہ" (həfʈɑ: vɑ:r əxbɑ:r rəsɑ:lɑ:/ daily or weekly publication).

To define the correct meaning of a word in its respective context is called lexical disambiguation or Word Sense Disambiguation (WSD) in the field of computational linguistics. WSD is defined as "the process of computationally determining which "sense" of a word is triggered by the use of the word in a particular context" [2]. The meaning of a word is determined by its usage hence unique meanings would arise with new usage patterns. To analyze this usage pattern, corpora need to be sense tagged.

A sense-tagged corpus is a significant linguistic resource because of the presence of semantic knowledge that is used in theoretical linguistics. In the field of computational linguistics, these resources are critically used for natural language processing (NLP) [3]. Moreover, statistics extracted from analysis of sense tagged corpus is used in a number of other research domains i.e. to extract and retrieve information, to summarize texts and to answer questions automatically [4].

There are three distinct approaches for word sense disambiguation. These approaches include knowledge based approach, unsupervised approach and supervised/semi-supervised approach [1]. Knowledge based approach uses methods that depend upon dictionaries or thesauri, but do not employ any corpus evidences. Unsupervised approaches include methods that depend upon un-annotated corpora. Supervised approach includes methods that use annotated corpora with sense IDs that act as seed data to train a WSD system. Therefore presence of an extensive sense tagged corpus is critical for successful WSD program.

This paper illustrates the development of an Urdu Sense Tagged Corpus Ver. 1.0. The final corpus consists of 5,611 sentences with 100K words of which 17,006 words are sense tagged. The paper is organized as follows. Section 2 reviews various manually sense tagged corpora constructed using respective language's WordNet. Section 3 then specifically details the development of Urdu sense tagged corpus, by first presenting the lexical resources used and the process followed in the senses annotation and Section 4 presents the current state of Urdu sense tagged corpus. Lastly, section 5 and 6 describes the research

[1] http://wordnet.princeton.edu

challenges in the development of Urdu sense tagged corpus and future work in this context.

## 2. Literature Review

In a comprehensive survey conducted by Bond [5], almost all available WordNet tagged corpora (along with their availability) for English and for other languages have been enlisted. Among these sense tagged corpora for English language, HECTOR, SemCor and DSO corpus have been explained below in further details. HECTOR was a collaborative project by Oxford University Press and Digital Education project [6] in corpus lexicography. This corpus consists of 20 million words (a pilot for the British National Corpus). A dictioary was developed alongside the tagging process and its senses were used to tag corpus instances [6].

The English SemCor corpus is a sense-tagged corpus of English [7], created very early in the WordNet project and is one of the first sense-tagged corpora produced for any language. The corpus consists of a subset of the Brown Corpus (700,000 words, with more than 200,000 sense-annotated) and it has been part-of-speech-tagged and sense-tagged. It is distributed under the Princeton WordNet License.

The DSO corpus includes 192,800 instances of frequently used nouns (121) and verbs (70) of English [6]. These occurrences have been hand tagged with WordNet 1.5 senses [8]. It is distributed on the Linguistic Data Consortium Catalogue[2] (LDC) under different licenses for LDC Members (free for 1997 members) and non-members. Unlike Semcor, which adopted all-word corpus approach i.e. tagging all words in the form of running text and therefore making it difficult for the supervised system to learn as there are inadequate number of examples per word, the DSO corpus uses the targeted tagging approach i.e. tagging all occurrences of a target word. The approaches used for DSO corpus resulted in the consequent evaluation efforts of Senseval [6].

Sense tagged corpora have been developed for other languages as well. For example, Japanese SemCor (JSEMCOR) [9] has been constructed using annotation transfer approach. According to this approach, sense tagged corpus in one language is translated into the target language and sense annotations are also projected to the target language. The sense projection is carried out using a WordNet in the target language which is aligned with the WordNet that was used to sense tag the source language text. The source corpus used is English SemCor and the source WordNet is Princeton (1.6) WordNet of English. The target language WordNet is Japanese WordNet [9]. The final corpus consists of 14,169 sentences with 150,555 content words of which 58,265 are sense tagged. The license is similar to the Princeton WordNet License, so the data is freely available.

Attempts have been made via DutchSemCor project [10] to develop a Dutch corpus. The instances in this corpus have been assigned senses from Cornetto lexical database [11] with a semi-automatic approach. In DutchSemCor about 282,503 tokens) have been manually tagged by two annotators (more than 400,000 have been manually tagged by at least one annotator and millions have been automatically tagged). This corpus was built in two phases. First phase deals with manual collection of 25 examples for each sense which were then used to the train a supervised WSD system. The system then looks for other 75 examples. The technique used for guiding the system to select suitable examples was active learning [12]. Dutch-SemCor is not available, but excerpts and statistics are freely downloadable.

Similarly, Bulgarian word sense tagged corpus[3] [13] has been constructed from Bulgarian Brown corpus. It consists of 811 excerpts each containing 100+words: the total size of the source corpus is 101,062 tokens. The words in BulSemCor are assigned meaning manually from the Bulgarian WordNet [14]. The sense-annotated corpus consists of 99,480 lexical units annotated with the most appropriate synset from the Bulgarian WordNet (BulNet). The corpus excerpts are offered under MS No Redistribution Non Commercial license for free, it is also possible to query the corpus online. The restrictions on use and redistribution mean that corpus is not considered open source.

An attempt for devising word sense annotated corpus for Chinese language has been made [15]. The subsequent work contains three components; a corpus, a lexicon and the linking between the lexicon and the Chinese Dictionary. The lexicon includes the description of 813 nouns and 132 verbs and 60,895 word instances have been tagged. This corpus is taken from three-month texts of People's Daily, an official daily newspaper for the government of China, with a move to extend this corpus for other kinds of texts.

To develop a successful and accurate WSD extensive sense tagged corpora are required [16]. It is argued that no essential growth can be made in the field of WSD until extensive lexical resources are built [17]. From the above reviewed literature, it is evident that the number of sense tagged corpus for English and for other languages increased in the past years. In the field of Urdu lexical resource development, some

[2] See catalog.ldc.upenn.edu/LDC97T12

[3] See dcl.bas.bg/en/corpora_en.html#SemC

attention has been paid to corpus construction, POS tagging and WordNet development by Center for Language Engineering, by making the CLE Urdu Digest Corpus[4] [18] and Urdu WordNet [19] available for research, but significant research still needs to be focused on sense annotation.

## 3. Developing Urdu Sense Tagged Corpus

This paper describes the development of an Urdu corpus annotated with word senses, in order to build a comprehensive resource for Urdu lexical semantics. In the construction of sense tagged corpus, four linguistic resources i.e. Urdu Wordnet 1.0 Wordlist, CLE Urdu Digest Corpus, Urdu WordNet and Urdu Morphological Analyzer have been used. The detail of these resources along with the explanation of annotation method and annotation tool is described in the following sections.

### 3.1. Linguistic Resources and Applications

Fundamentally four major resources have been used to develop the Urdu Senses tagged corpus; Urdu Wordnet 1.0 Wordlis, CLE Urdu Digest Corpus, Urdu WordNet and Urdu Morphological Analyzer.

#### 3.1.1. Urdu WordNet 1.0 Wordlist

The Urdu WordNet 1.0 Wordlist[5] used in sense tagging comprises 5000 words of which approximately 3000 words have been taken from the following three sources, 18 million words corpus extracted from online news websites, CLE Urdu Digest Corpus and Urdu Verblist[6] extracted from Online Urdu Dictionary (OUD[7]). Additional 2000 words have been included alongside the process of Urdu WordNet development.

#### 3.1.2. CLE Urdu Digest Corpus

The initial data for sense tagging has been taken from CLE Urdu Digest Corpus 100K [18] containing 102,209 words of Urdu. This corpus covers multiple domains and genres and is designed with a move to be used in linguistic research.

CLE Urdu Digest Corpus consists of two main categories i.e. Informational and Imaginative. Informational domain covers 80% of the corpus while imaginative domain covers 20% of the corpus. The data for this corpus construction has been taken from Urdu Digest[8]. The data used in this corpus ranges between years 2003- 2011. The data is genre wise distributed. There are 348 files saved in UTF-8 format and each files includes 300 words approximately [18].

This Corpus has been manually tagged with parts of speech, using the revised POS tagset [20]. 80% of the corpus was then used to train the Urdu POS tagger available on CLE website. The tagger was then tested on 20% of the corpus. The files for POS tagging were selected randomly. The results showed tagging accuracy of 96.8%.

#### 3.1.3. Urdu WordNet

Urdu WordNet[9] is a semantic dictionary of Urdu [19] developed by Center for Language Engineering. It contains 5058 senses approximately. All synsets have POS definition, unique synset ID, definition, synset and example. The example of an entry has been given in Figure 1.

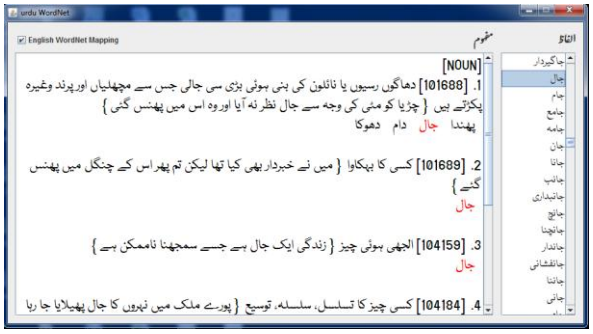


**Figure 1 Layout of Urdu WordNet**

#### 3.1.4. Urdu Morphological Analyzer

An Urdu Morphological Analyzer [21] is used for showing all the possible morphological forms of words. This analyzer is then integrated in the sense tagging tool displaying corpus matches for a certain word being sense tagged, in order to display all possible morphological forms of a base word along with its all base-form occurrences. This functionality provides additional data to an annotator following the target annotation approach, through which in addition to the specific word entry, specific entries of various morphological forms of the word are also displayed to the annotator for sense tagging.

[4] http://cle.org.pk/clestore/index.htm
[5] http://www.cle.org.pk/software/ling_resources/UrduWordNetWordlist.htm
[6] http://www.cle.org.pk/software/ling_resources/urduverblist.htm
[7] http://182.180.102.251:8081/oud/default.aspx
[8] www.urdudigest.pk
[9] http://www.cle.net.pk/urduwordnet/

## 3.2. Sense Annotation Method and Tool

The following sub-sections give detail about the method used for sense tagging and Urdu word sense annotation tool i.e. ʊrd̪u məʃhu:m kɑ:r / اردو مفہوم کار.

### 3.2.1. Sense Annotation Method

The word forms in the corpus were POS-tagged and linked to the corresponding word senses in Urdu WordNet, if available. Conventionally, two annotation methods are used in corpus sense tagging. Firstly the sequential tagging method in which the corpus text appears in the form of a running text and the words are tagged in a sequence as they appear. The second approach is the targeted tagging method in which all instances of the target word appear with complete pre and post context of the word and the annotator is facilitated to make contrasts of the contexts. The targeted tagging approach enables the annotator to review all occurrences of the target word with its meanings only once and therefore implement a more consistent comparison of the different contexts. On the other hand, sequential tagging approach requires the annotator to alter attention all the time to multiple words as they appear in the text, followed by extensive cognitive load. Considering the mentioned advantage of using the former approach, CLE Urdu Sense tagged Corpus Ver. 1.0 has been annotated using the targeted tagging approach.

### 3.2.2. Urdu Word-Sense Annotation tool (ʊrd̪u məʃhu:m kɑ:r / اردو مفہوم کار)

For the purpose of word sense annotation, a word sense annotation tool has been developed by Center for Language Engineering, which enables manual disambiguation of large volume of texts. This annotation tool uses POS tagged files as input and generates sense tagged files (where the words are tagged with synset ID) as output using the Urdu synset ID developed through the Urdu WordNet. The user interface of the tool gives three views presented in the 1, 2 and 3 numbered windows in figure 2.

#### 3.2.2.1. Selection view

The selection view of the interface displays the list of high frequency words and enables the annotator to select a target word. This window makes use of Urdu wordlist (discussed in section 3.1.1.). The tool then matches these words with those in the corpus and WordNet.

#### 3.2.2.2. WordNet view

This view displays the linguistic information available in WordNet for the selected lexical item. This window integrates file generated via Urdu WordNet and enables the reader to select the most appropriate sense among the available senses and the POS by matching it with the POS used in the corpus. With the help of this view, the annotator can also compare the contexts of a sense by its use in the WordNet example and its occurrence in the corpus.

#### 3.2.2.3. Corpus view

The corpus view of the interface displays the corpus with all occurrences of the target word. Facilitated by the integration of the Morphological Analyzer, this corpus view not only displays all occurrence of the specific word in the corpus, but additionally displays all occurrences of the word's complete morphological forms available in the Corpus. As this window shows the complete sentence within which the target word occurs thus, the annotator is facilitated to comprehend the complete pre and post context of the word occurrence, in order to mark the most appropriate sense from the available word senses.

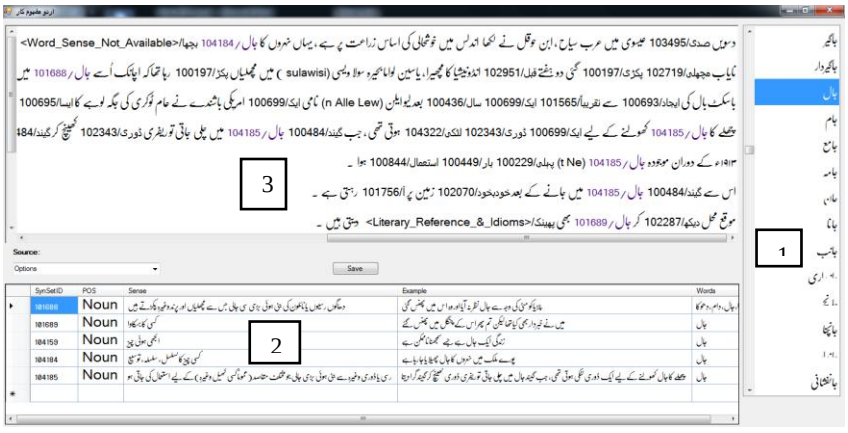


**Figure 2 Layout of the annotation tool showing word sense annotation**

### 3.2.3. Corpus Annotation Tags

The annotator selects a specific word from the selection view and its complete occurrence and occurrences of its morphological forms are highlighted (in a new color) in the Corpus view window. Based on the available different senses of the selected (target) word in the WordNet view, the annotator carefully tags every occurrence. Ideally, if the specific sense of a word occurrence exactly matches with the sense entered in the WordNet, the annotator selects that sense and the respective occurrence gets tagged with that sense ID. If the sense of a specific target word is not available in the WordNet view, then the annotator reports the problem in four possible options, provided in the annotation utility. These are also displayed in Figure 3 below.

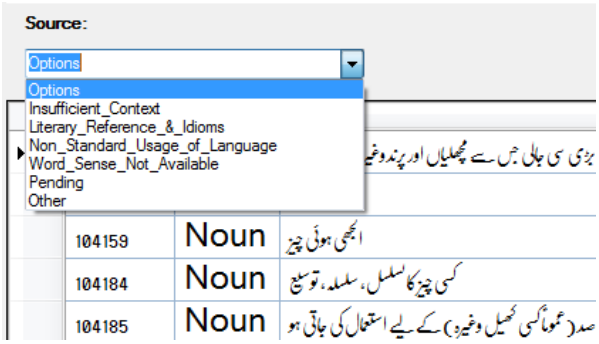


**Figure 3 Snapshot of the annotation tool showing tagging options**

#### 3.2.3.1. Insufficient Context

This is the case where some contexts were too brief for a sense to be assigned to the word e.g. in the following sentence the context of the word u:pər (اوپر/over) is too brief to be sense tagged.

ہی <Insufficient_Context اس طرح اوپر/>
- کئی راستے بن گئے <Insufficient_Context>/اوپر

ɪs ʈərəh u:pər hi: u:pər kəi: rɑ:sʈe: bən gəe:
this like over and over many ways made were
Like this, many ways were made.

#### 3.2.3.2. Literary Reference/symbolic Sense

This tag is used to tag idiomatic phrase in corpus or the case where a corpus sense is not conveying its literal meaning rather it has been used in symbolic sense e.g. in the following sentence nəbzẽ: (نبضیں/pulse) cannot be sense tagged as it has been used as an idiom here.

دھیمی <Literary_Reference_&_Idioms>/ وقت کی نبضیں چل رہی تھیں۔

vəkʈ ki: nəbzẽ: ɖʰi:mĩ: ʧəl rəhi: ʈhĩ:
time slowly passing was
Time was passing slowly.

#### 3.2.3.3. Non-standard Usage of language

In context where people use the semantic aspects of language imaginatively and creatively in the same way as they do not always follow the rules of grammar and syntax, this tag is used. Apart from some recognized figurative sense extensions in the dictionary, this aspect of language use is unpredictable.

#### 3.2.3.4. Word Sense Not Available

This is the context where the particular sense of the word being displayed in the corpus view is valid but not available in the WordNet view. This option was specifically designed to provide feedback to the Urdu WordNet development team for inclusion of missing senses for a word.

#### 3.2.3.5. Other

This is the category left for the addition of any other comment by the annotators and is tagged for discussion and mutual agreement.

## 4. Current status of Urdu sense tagged corpus

The corpus used for sense tagging is POS tagged CLE Urdu Digest Corpus [20]. Sense layer has been added to this corpus manually over a period of ten months by a single annotator using sense annotation tool ʊrɖu məfhu:m kɑ:r/اردو مفہوم کار described in section 3 above. The current status of Urdu sense tagged corpus has been given in the table 1 and 2 below. Table 1 shows that the final corpus consists of 5611 sentences with 100K words of which 17006 are sense tagged.

**Table 1: Current state of Urdu sense tagged corpus**

| Sense tagged corpus | |
|---|---|
| Total no. of sentences in the corpus | 5611 |
| Total no. of words in the corpus | 100,000 |
| Tagged total word types | 2225 |
| Tagged total sense types | 2285 |
| Tagged total word tokens | 17006 |

Table 2 shows that there are 559 words which have more than 2 senses tagged and 1522 words have one sense tagged in the corpus.

**Table 2: Word count with no. of senses tagged**

| | No. of senses tagged | | | | | |
|---|---|---|---|---|---|---|
| **Words** | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1522 | 345 | 118 | 49 | 21 | 14 |
| | **No. of senses tagged** | | | | | |
| **Words** | 7 | 8 | 9 | 10 | 11 | 12 |
| | 3 | 2 | 3 | 3 | 1 | 1 |

The work is proceeding to add more examples per sense to aid the development of an automatic sense tagger.

## 5. Improving WordNet via sense tagging process

During the course of annotation, the annotators followed certain consistency criteria which helped in improving WordNet as well. Following identifications were made during the process of tagging:

### 5.1. Consistency of sense definition with POS

In some cases, it was observed that the interpretative definition (gloss) associated with the synset doesn't match with POS of the word e.g. initially the sense of word əfsurɖa: (افسردہ/sad) was more like a verb than an adjective. After feedback to the WordNet team, it was then modified to convey adjective sense.

**Table 3: Consistency of Definition with POS**

| Words | POS | Sense | Modified |
|---|---|---|---|
| افسردہ | ADJ | تھکن یا غم اور دکھ ہونے کی وجہ سے بیزار ہوجانا | تھکن یا غم اور دکھ ہونے کی وجہ سے بیزار ہونے والا، بجھا بجھا سا |

### 5.2. Consistency of sense example with POS

In some cases, it was observed that the example associated with the synset doesn't matchwith POS of the word e.g.

**Table 4: Consistency of Example with POS**

| Word | POS | Sense | Example | Modified |
|---|---|---|---|---|
| خزندہ | Noun | رینگنے والاجانور | وہاں خزندہ جانوروں کی بھرمار تھی | وہاں خزندوں کی بھرمار تھی |

### 5.3. Consistency of definition across synset

As the definition associated with the synset encodes the meaning of all the members of the synset in an explicit way, it is very important that all the members have equal relationship with sense meaning. Substitution tests were applied to identify semantic equivalents of words found in the corpus and only those synonyms were pertained which have equal chance of usage in the example sentence e.g. ʊɖʱe:rnɑ: (ادھیڑنا) cannot be the part of synset in this particular sense.

**Table 5: Consistency of definition across synset with POS**

| POS | Sense | Example | Words | Modified |
|---|---|---|---|---|
| Verb | پرت والی چیز کا کوئی حصہ الگ کرنا، ٹکڑے کرنا | اس نے تختہ توڑ کے کشتی کو پھاڑ دیا | پھاڑنا، چیرنا ادھیڑنا | پھاڑنا،چیرنا |

### 5.4. Addition of senses available in the corpus

In the development of WordNet, less frequent senses were not added. During the process of annotation all the "word sense not available" tags were reported back to the WordNet team so that those particular senses can be added in the WordNet according to their usage in the corpus.

## 6. Challenges in the Process of sense tagging

Annotators faced two specific challenges during the process of tagging; a) tool specific limitation i.e. in some cases, annotator tool couldn't match the corpus instances b) language specific limitations i.e. annotators faced difficulty in tagging certain linguistic contexts such as; non-standardized translations, foreign language borrowed words and complex predicates. The detail of these ambiguous contexts is given below:

### 6.1. Non-standardized translations

It was ambiguous to tag non-standardized translations of English words which have become part of Urdu language e.g. sense mapping was not found for

bʊlənɖ fiʃare xun (بلند فشارِ خون) i.e. high blood pressure and ʃəmsi ʈəxʈe (شمسی تختے) i.e. solar panels.

### 6.2. Foreign language borrowed words

The sense mapping was not found in the dictionaries for those words which are borrowed from foreign language and have been lexicalized for Urdu e.g. test match, basket ball and interview.

### 6.3. Complex Predicates

A fundamental assumption underlying all syntactic theories has been that the main verb plays the role of predication within a clause and all other elements in the clause are either arguments or modifiers [22]. But in Urdu there are complex predicates, defined as containing two or more predicational elements which jointly predicate within a mono-clausal structure. Annotators faced difficulty in tagging complex predicates i.e. the combinations of main verb and light verb for example in the following sentence the word paːeː ʤaːneː (پائےجانے) is a sense tagging challenge as this sense was not found in the available senses of word paːnaː (پانا) in the dictionary.

اس میں پائےجانے والی حیاتین ہماری آنکھوں کو صحت مند رکھتی ہے ۔

ɪs mẽː paːeː ʤaːneː vaːliː həjaːʈiːn həmaːriː ãːnkʰõː
this in found protein our eyes
koː sehəʈmənɖ rəkʰʈiː hæː
healthy keeps
The protein found in this keeps our eyes healthy.

### 6.4. Normalization

The annotation tool was unable to match corpus in some contexts e.g. vow (و) with hamza above case (ۥ). The reason is that these combinations were typed in different formats in corpus and WordNet and hence requiring the process of normalization [23] which is the process of representing texts into consistent formats.

## 7. Conclusion and Future Work

This paper describes the construction of a sense-tagged Urdu corpus. The goal of this research has been to create a valuable resource both for word sense disambiguation and researches on Urdu lexical semantics. The current corpus consists of 5611 sentences with 100K words of which 17006 words are sense tagged. This manually annotated corpus can act as a seed corpus for automated methods to extract additional senses and their relationships.

## 8. Acknowledgements

This work has been conducted through the Essential Urdu Linguistic Resources[10] project (www.cle.org.pk\eulr), supported through a research grant from DAAD, Germany

## 9. References

[1] A. Eneko and P. Edmonds, "*Word Sense Disambiguation: Algorithms and Applications*" Springer, 2007.

[2] A. Kilgarriff, "Word senses", *Word Sense Disambiguation*, Springer Netherlands, 2006.

[3] P. Resnik, "WSD in NLP applications", *Word Sense Disambiguation: Algorithms and Applications (2006).*

[4] S. J. Ker, C. R. Huang, J. F. Hong, S. Y. Liu, H. L. Jian, I. L. Su & S. K. Hsieh, "Design and Prototype of a Large-scale and Fully Sense-tagged Corpus." Large-Scale Knowledge Resources. Construction and Application. *Springer* Berlin Heidelberg, 2008.

[5] T. Petrolito and F. Bond, "A survey of WordNet annotated corpora", Proceedings of the *7th Global WordNet Conference* (GWC 2014), 2014.

[6] A. Kilgarriff, "SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Program", Proceedings of First International Conference on *Language Resources and Evaluation*, Granada, 1998.

[7] S. Landes, C. Leacock and R. I. Tengi, "Building semantic concordances", *WordNet: An electronic lexical database*, 1998: 199-216.

[8] M. Palmer, H. T. Ng and H. T. Dang, "Evaluation of WSD systems" *Word Sense Disambiguation: Algorithms and Applications*, 2006.

[9] F. Bond, T. Baldwin, R. Fothergill & K. Uchimoto, "Japanese SemCor: A sense-tagged corpus of Japanese", Proceedings of the 6th International Conference of the *Global WordNet Association* (GWC), 2012.

[10] P. Vossen, A. Görög, R. Izquierdo, & A. van den Bosch, "DutchSemCor: Targeting the ideal sense-tagged corpus", *LREC*, 2012.

[11] P. Vossen, I. Maks, R. Segers, & H. VanderVliet, "Integrating Lexical Units, Synsets, and Ontology in the Cornetto Database", Proceedings on the *sixth international Conference on Language Resources and Evaluation* (LREC), 2008.

[12] J. Zhu, H. Wang, T. Yao, & B. K. Tsou, "Active learning with sampling by uncertainty and density for word sense disambiguation and text

[10] See http://cle.org.pk/eulr/ .

classification", Proceedings of the *22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008.

[13] S. Koeva, S. Leseva, E. Tarpomanova, B. Rizov, T. Dimitrova & H. Kukova, "Bulgarian Sense-Annotated Corpus–Results and Achievements", *FASSBL7*, 2010: 41.

[14] S. Koeva, S. Mihov, & T. Tinchev, "Bulgarian Wordnet–Structure and Validation." *Romanian Journal of Information Science and Technology* 7, 2004.

[15] Y. Wu, P. Jin, Y. Zhang & S. Yu, "A Chinese corpus with word sense annotation" Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead, *Springer*, Berlin Heidelberg, 2006.

[16] Ng, H. T., "Getting serious about word sense disambiguation" In Proceedings of the *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, 1997.

[17] Véronis, J., "Sense tagging: does it make sense", In *Corpus linguistics conference*, 2001.

[18] S, Urooj, S, Hussain, F, Adeeba, F. Jabeen and R. Perveen, "CLE Urdu Digest Corpus", in the Proc. of *Conference on Language and Technology 2012* (CLT12), Lahore, Pakistan, 2012.

[19] A. Zafar, A. Mahmood, F. Abdullah, S. Zahid, S. Hussain, and A. Mustafa, "Developing Urdu WordNet Using the Merge Approach ", in the Proceedings of *Conference on Language and Technology* 2012 (CLT12), Lahore, Pakistan.

[20] T. Ahmed, , S. Urooj, , S. Hussain, A. Mustafa, R. Parveen, , F. Adeeba, A. Hautli, & M. Butt, "The CLE Urdu POS Tagset", poster presentation in *Language Resources and Evaluation Conference* (LERC 14), 2014, Reykjavik, Iceland.

[21] Hussain**,** S., "*Finite-State Morphological Analyzer for Urdu*", National University of Computer and Emerging Sciences, (2004), Lahore, Pakistan.

[22] M. Butt, "The light verb jungle: Still Hacking Away", in Workshop on Multi-Verb Constructions, 2003.

[23] S. Hussain, S. Gul and A. Waseem, "Developing lexicographic sorting: An example for Urdu", in *ACM Transactions on Asian Language Information Processing* (TALIP), Volume 6 Issue 3, 2007.